Prediction of Soil Particle Size Fraction using Geographically Weighted Regression and Random Forest

HENNY PRAMOEDYO[1], NOVI NUR AINI[1], SATIVANDI RIZA[2], DANANG ARIYANTO[3]
[1]Department of Statistics, University of Brawijaya
Veteran St, Malang, East Java
INDONESIA

[2]Department of Soil Science, University of Brawijaya
Veteran St, Malang, East Java
INDONESIA

[3]Department of Mathematics, University of Brawijaya
Veteran St, Malang, East Java
INDONESIA

Abstract: The development of spatial modeling for soil properties has progressed in recent decades. This responds to the growing demand for land spatial data and exact soil property prediction for agronomical reasons, particularly in precision farming, in order to speed up precision agricultural activities. In this regards a comparison of the GWR and RF models was carried out in order to determine which model is the best at forecasting surface soil texture and how dependable each model is at doing so. The purpose of this research is to get the best model in predicting particle soil fraction (PSF). 50 topsoil samples were collected from several locations in the Kalikonto Watershed, Indonesia, and the soil PSF (sand, silt, and clay) in the upper 10 cm varied. The LMV, slope, and elevation were calculated using DEM data and utilized as predictor variables. As a result, the weighting of the GWR model has a considerable impact on the final model, and all other factors have a major effect on the PSF determination. The RF, on the other hand, looks to be superior than the GWR variants. The RF model outperformed the other models in every PSF variable. This study reveals that topsoil quality and terrain attributes are linked, which may be assessed using field measurements and model projections. More research is needed to generate more efficient input parameters that will help with soil variability precision and accuracy of soil map products.

Key-Words: DEM, Particle Size Fraction, Modelling, Geographically Weighted Regression, Random Forest, Prediction

Received: June 30, 2021. Revised: November 20, 2021. Accepted: December 13, 2021. Published: December 27, 2021.

1 Introduction

In the recent few decades, the development of spatial modeling for soil properties has advanced. This addresses the increased need for land spatial data information and precise soil property predictions for agronomical reasons, notably in precision farming, in order to accelerate precision agricultural activities [1]. Besides that soil modeling is also very important for foundation modelling [2]. One of the soil properties that cannot be overlooked is soil texture. The flow of water, heat, and nutrients, as well as the form and stability of the soil structure, are all influenced by soil texture and, of course, particle size distribution. It is required to model soil texture since it is a compositional data set that specifies the particle size of the soil mineral fraction with the variables of sand, silt, and clay [3] [4].

The use of statistical approaches and geomorphology to correlate with landscape features produced from a digital elevation model (DEM) and remote sensing data tends to support spatial prediction of soil properties [3], [4]. Topographic variability can be determined using DEM data, and used as a predictor or independent variable in forecasting soil texture. Multiple Linear Regression is a widely used method for modeling or predicting an item by examining the relationship between the dependent variable and a set of independent factors as well as Regression Tress [1], [5], [6]. Other statistical modeling techniques, such as Geographically Weighted Regression (GWR) are

less popular for modelling the soil PSF. Therefore, the GWR technique was used in this study, which was combined with topographic variables and then compared to Random Forest (RF) approach, which has formerly been used in PSF modeling studies.

The GWR model is a regression approach that provides locally linear regression estimators for each data point or location [7]. This method is an extension of linear regression analysis that takes into account spatial dimension [8]. Response variables are predicted using predictor variables, the regression coefficients of which are dependent on the location of the data [8], [9]. The GWR method is not the same as the RF method, which is based on the creation of a single Decision Tree method.

Multiple trees are used in the RF approach, each of which is trained on a set of sample data. The class is determined by the amount of votes received from each tree. The ensemble learning methodology is more accurate than other machine learning approaches because it uses a mixture of several classifiers to provide more accurate results than a single classifier [10], [11]. The ability to model highly nonlinear dimensional relationships, the use of categorical and continuous variables, resistance to "overfitting," relative robustness in the presence of data noise, the establishment of an impartial measure of the error rate, the ability to determine the relevance of the variables used, and the requirement for few parameters for implementation are just some of the benefits [12].

The objectives of this paper is to compare the GWR and Random Forest models to find a best modelling in order to achieve high-accuracy prediction model for PSF projections. The two methods will be compared to see which is the most accurate model for predicting surface soil texture and how trustworthy each model is. It can be used as a reference in management to improve land use if there is a credible instrument for predicting surface soil texture.

2 Problem Formulation

2.1. Study Site

The research study was conducted in the Kalikonto Watershed in East Java, Indonesia. The map of the area of study is shown in Figure 1.

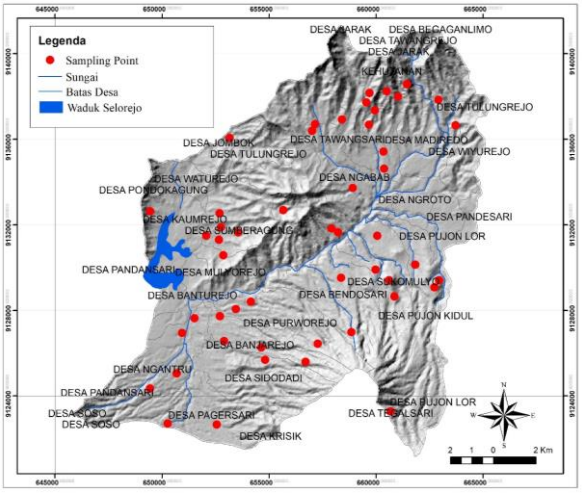

Fig 1. Location map of the study area

This region is made up of the inter-volcanic plains between the Anjasmara Tua Mountain in the north and the Butak-Kawi Mountain in the south. The majority of the land in the study area was in agricultural regions. The physiography of the area is made up of 235.7 km^2 of undulating hills and plains. 50 topsoil samples were collected at diverse sites to determine the topsoil layer quality, and these samples had varying soil PSF (sand, silt, and clay) in the topmost 10 cm.

2.2. Data Set

Based on DEM data, the LMV, slope, and elevation were computed to be used as predictor variables. The DEM data served as the analyses' principal input. For the entire watershed, we took 30 m SRTM (Shuttle Radar Topography Mission) DEM data from the USGS data source to get topographic variables. Slope, Elevation, and 6 Local Morphologic Factors (LMV) that revealed the diversity of the curvature of a topography [13] were the variables in this investigation:

1. Vertical Curvature (Kv)

$$K_v = \frac{p^2r + 2pqs + q^2t}{(p^2+q^2)\sqrt{(1+p^2+q^2)^3}}$$

2. Horizontal Curvature (Kh)

$$K_h = \frac{q^2r - 2pqs + p^2t}{(p^2+q^2)\sqrt{1+p^2+q^2}}$$

3. Accumulation Curvature (Ka)

$$K_a = \frac{(q^2r - 2pqs + p^2t)(p^2r + 2pqs + q^2t)}{[(p^2+q^2)(1+p^2+q^2)]^2}$$

4. Ring Curvature (Kr)

$$K_r = \left[\frac{(p^2-q^2)s - pq(r-t)}{(p^2+q^2)(1+p^2+q^2)}\right]^2$$

5. Northness Aspects (An)

$$A_n = cos\left[\begin{matrix}-90[1-sin(q)](1-|sin(p)|)+\\ 180[1+sin(p)]-\frac{180}{\pi}\,sin(p)arcos\left(\frac{-q}{\sqrt{p^2+q^2}}\right)\end{matrix}\right]$$

6. Eastness Aspects (Ae)

$$A_e = sin\left[\begin{matrix}-90[1-sin(q)](1-|sin(p)|)+\\ 180[1+sin(p)]-\frac{180}{\pi}\,sin(p)arcos\left(\frac{-q}{\sqrt{p^2+q^2}}\right)\end{matrix}\right]$$

However, in order to acquire these variables, an analysis of the DEM data must first be performed in order to obtain the derivative value of the elevation, which is the DEM data's digital number value. The following formula is used to calculate the elevation derivative value:

$$p = \frac{z_3 + z_6 + z_9 - z_1 - z_4 + z_7)}{6w^2}$$

$$q = \frac{z_1 + z_2 + z_3 - z_7 - z_8 - z_9)}{6w^2}$$

$$r = \frac{z_1 + z_3 + z_4 + z_6 + z_7 + z_9 - 2(z_2 + z_5 + z_8)}{3w^2}$$

$$s = \frac{z_3 + z_7 - z_1 - z_9}{4w^2}$$

$$t = \frac{z_1 + z_2 + z_3 + z_7 + z_8 + z_8 - 2(z_4 + z_5 + z_6)}{3w^2}$$

The elevation is z, and the cell size is w in pixels [14]. To obtain the z value, a measuring window must be used, as shown:

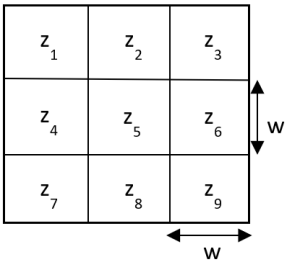

Fig. 2: Illustration of the measurement window to get the elevation derivative value (p, q, r, s and t).

2.3. Statistical Analysis

2.3.1. Geographically Weighted Regression (GWR)

Geographically weighted regression is a localized version of classic multiple linear regression, in which regression coefficients are particular to an area rather than worldwide estimates. A basic GWR model's specification is as follows:

$$y_i = \beta_0(u_i,v_i) + \sum_{k=1}^{p} \beta_k(u_i,v_i)x_{ik} + \varepsilon_i\,; \quad i = 1,2,\dots,n \tag{1}$$

where y_i is the dependent variable at location i, x_{ik} is the value of the kth explanatory variable at the i location, the $\beta_k(u_i,v_i)x_{ik}$ is the local regression coefficient for the kth explanatory variable at location i, $\beta_0(u_i,v_i)$ is the intercept parameter at location i, and ε_i is the random disturbance at location i, which may follow an independent normal distribution with zero mean and homogeneous variance[9]. The GWR model can be expressed in matrix notation to make it easier to understand:

$$\boldsymbol{Y_w = X_w\beta_w + \varepsilon_w} \tag{2}$$

For parameter estimation in Geographically Weighted Regression, Weighted Least Square (WLS) is utilized, so:

$$\begin{aligned} L &= \boldsymbol{\varepsilon_w}^T\boldsymbol{\varepsilon_w} \\ &= (\boldsymbol{Y_w - X_w\beta_w})^T(\boldsymbol{Y_w - X_w\beta_w}) \\ &= \boldsymbol{Y_w}^T\boldsymbol{Y_w} - \boldsymbol{\beta_w}^T\boldsymbol{X_w}^T\boldsymbol{Y_w} - \boldsymbol{Y_w}^T\boldsymbol{X_w\beta_w} + \boldsymbol{\beta_w}^T\boldsymbol{X_w}^T\boldsymbol{X_w\beta_w} \\ &= \boldsymbol{Y_w}^T\boldsymbol{Y_w} - 2\boldsymbol{\beta_w}^T\boldsymbol{X_w}^T\boldsymbol{Y_w} + \boldsymbol{\beta_w}^T\boldsymbol{X_w}^T\boldsymbol{X_w\beta_w} \end{aligned} \tag{3}$$

A univariate optimization, as we learnt in calculus, entails taking the derivative and putting it equal to 0. This provides us with,

$$\frac{\partial L}{\partial \boldsymbol{\beta_w}} = \frac{\partial(\boldsymbol{Y_w}^T\boldsymbol{Y_w} - 2\boldsymbol{\beta_w}^T\boldsymbol{X_w}^T\boldsymbol{Y_w} + \boldsymbol{\beta_w}^T\boldsymbol{X_w}^T\boldsymbol{X_w\beta_w})}{\partial \boldsymbol{\beta_w}} = 0$$

$$\frac{\partial L}{\partial \boldsymbol{\beta_w}} = -2\boldsymbol{X_w}^T\boldsymbol{Y_w} + 2\boldsymbol{X_w}^T\boldsymbol{X_w}\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}} = 0$$

$$\frac{\partial L}{\partial \beta_w} = -X_w{}^T Y_w + X_w{}^T X_w \hat{\beta}_w = 0$$

$$X_w{}^T X_w \hat{\beta}_w = X_w{}^T Y_w$$

$$\hat{\beta}_w = (X_w{}^T X_w)^{-1} X_w{}^T Y_w \quad (4)$$

If $W^{\frac{1}{2}}Y = W^{\frac{1}{2}}X\beta + W^{\frac{1}{2}}\varepsilon$ is equal to $Y_w = X_w\beta_w + \varepsilon_w$ so,

$$= \left[(W^{\frac{1}{2}}X)^T W^{\frac{1}{2}}X\right]^{-1}(W^{\frac{1}{2}}X)^T W^{\frac{1}{2}}Y$$

$$= (X^T W^{\frac{1}{2}} W^{\frac{1}{2}}X)^{-1} X^T W^{\frac{1}{2}} W^{\frac{1}{2}} Y$$

$$\hat{\beta}_w = (X^T W X)^{-1} X^T W Y \quad (5)$$

The Geographically Weighted Regression model parameter estimation is achieved by matrix operation for each i-th point:

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i)\, X)^{-1} X^T W(u_i, v_i)\, Y$$

$$i = 1,2,\dots,n \quad (6)$$

with,

$$W_{ij}(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}$$

The spatial weighting matrix, which can be constructed in a variety of ways, is the initial stage in estimating parameters in GWR. W*ij* can be specified as a continuous and monotonic decreasing function of the distance dij between points *i* and *j*, for example. The weight of each point can be determined using the Gaussian function [15] for adaptive kernel size:

$$w_{ij}(u_i, v_i) = exp\left(-\left(\frac{d_{ij}}{h_{i(q)}}\right)^2\right) \quad (7)$$

And the Bisquare Function follow as :

$$w_{ij}(u_i, v_i) \begin{cases} 1 - \left(\frac{d_{ij}}{h_{i(q)}}\right) & ,if\ d_{ij} < h_i \\ 0 & ,if\ d_{ij} > h_i \end{cases} \quad (8)$$

where $w_{ij}(u_i, v_i)$ is the weight of position j in the space where data are seen for estimating the dependent variable at location *i*, and h_i is referred as a bandwidth. d_{ij} is eucledian distance between points i and j, $d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$ and $h_{i(q)}$ is the adaptive bandwidth that sets q as the nestest neighbor from i location for each location. The term bandwidth is used to describe how the kernel's size should be determined. It determines how smooth the model will be. Cross-validation (CV) is an iterative procedure for finding the kernel bandwidth that minimizes the prediction error of all *y*(*s*) using only a fraction of the data[16].

$$CV(h) = \sum_{i=1}^{n}(y_i - \hat{y}_{\neq i}(h))^2 \quad (9)$$

where $\hat{y}_{\neq i}(h)$ is the predicted value of observation *i* with calibration location *i* left out of the estimation dataset.

The test for spatial autocorrelation is the next stage. For spatial autocorrelation, Moran's I is a well-known test. Covariance and correlation statistics are analogous to the index. The product of the divergence between each value and the estimate of the global mean x is used to calculate the degree of similarity between values at two locations I and j. The sum of the resulting values for all pairs of places is the spatial autocovariance, which is weighted by their spatial proximity. The standardized index is expressed in the following way: $I = \frac{N}{S_0}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (10)

Where :

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}$$

The test of GWR's parameters model partially using t test with hypothesis as follows [1]:

$$H_0 : \beta_j(u_i, v_i) = 0$$

$$H_1 : \beta_j(u_i, v_i) \neq 0\ ;\ j = 1,2,\dots,k$$

t test statistic can be written as:

$$\frac{\hat{\beta}_k(u_i v_i)}{\hat{\sigma}\sqrt{c_{jj}}} \sim t_{(n-k-1)} \quad (11)$$

Where c_{jj} is a diagonal element of the $\mathbf{CC^T}$ matrix,

with $\boldsymbol{C} = \left(\boldsymbol{X^T W(u_i,v_i)X}\right)^{-1}\boldsymbol{X^T W(u_i,v_i)}$.

The other step is testing parameter of GWR model simultanously, the hypothesis is[9]:
H_0 : $\beta_j(u_i,v_i) = \beta_j$, where j=1, 2, .. k
H_1 : at least one $\beta_j(u_i,v_i)$ has a relation with location (u_i,v_i)
The statistic test is:

$$\frac{SSE(H_1)/\left[\frac{\delta_1^2}{\delta_2}\right]}{SSE(H_0)/n-k-1} \sim F^*_{\left(\frac{\delta_1^2}{\delta_2},n-k-1\right)} \quad (12)$$

where:

$$SSE(H_0) = \boldsymbol{y^T(I-L)^T(I-L)y}$$

$$SSE(H_1) = \boldsymbol{y^T(I-S)y}$$

$$\delta_1 = trace\{\boldsymbol{(I-L)^T(I-L)}\}$$

$$\delta_2 = trace\{\boldsymbol{(I-L)^T(I-L)}\}^2$$

$$\boldsymbol{L}_{(n\times n)} = \begin{pmatrix} x_1^T[X^TW(u_1,v_1)X]^{-1}X^TW(u_1,v_1) \\ x_2^T[X^TW(u_2,v_2)X]^{-1}X^TW(u_2,v_2) \\ \vdots \\ x_n^T[X^TW(u_n,v_n)X]^{-1}X^TW(u_n,v_n) \end{pmatrix}$$

$$\boldsymbol{S} = \mathbf{X(X^TX)^{-1}X^T}$$

$\mathbf{I}$ = identity matrix ordo n

The initial stage in this study is to use Moran's I [17] to do a spatial autocorrelation test. The next step is to determine the Euclidean distance between two points of observation using longitude and latitude as the coordinates (1). Then Using Cross Validation (CV) [18], determine the best bandwidth (h) for all observation locations. With the Adaptive Gaussian Kernel and Adaptive Bisquare Kernel weighting functions [18], it calculates a weighted matrix by entering the Euclidean distance and the optimum bandwidth of each technique.

The coefficient of determination (R2) was used to assess the model's performance. The next step is to compare the two models based on the coefficient of determination (R2) and Root Mean Square Error to determine which is the best (RMSE).
The size of the difference between the actual value and the expected result value can be determined using the RMSE function. The following formula calculates the RMSE value:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_i - \hat{Y}_i)^2} \quad (13)$$

The coefficient of determination (R^2) stating how much diversity of dependent variables can be explained by independent variables. The value R^2 is derived from the following formula:

$$R^2(u_i,v_i) = \frac{JKR_w}{JKT_w} = \frac{\sum_{j=1}^{p} W_{ij}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{p} W_{ij}(y_j - \bar{y})^2} \quad (14)$$

The next stage is to estimate the GWR model's parameters, then test them simultaneously with the F test (12) and partially with the t test (11). We created a map based on the data after gathering predictor factors with a substantial influence.

2.3.2 Random Forest

Random forest [19] is a classification that consists of many decision trees constructed from random vectors. In the classification process, the individual is based on the vote of the most votes in the group population tree [20]. Random forest is the development of a decision tree by using several decision trees where each decision tree has been trained using individual samples and each attribute is divided into a tree that is selected between a subset of random attributes and each attribute is divided into a tree that is selected between a subset of random attributes and each attribute is divided into a tree that is selected between a subset of random attributes and each attribute Random forest is a randomization technique that gathers independent variables as well as sample data, resulting in a classification tree of varied sizes and shapes [21].

The random forest operator produces a collection of random trees, with the class created by the classification process being chosen from the most classes (mode) generated by the current random tree [10]. In the random forest approach, many trees are produced, resulting in a forest that will be examined. Random forest is applied to a data cluster with n observations and n explanatory factors by [11]:

1. Perform n-fold random sampling with recovery on data clusters, with this step serving as the bootstrap stage.
2. The tree is built until it reaches its maximum size using the bootstrap example (without pruning).

The best sorter is determined based on these m explanatory variables at each node, where m << p. This stage is known as the random feature selection stage.

3. To make a forest with k trees, repeat steps 1 and 2 k times.

The random forest method must determine m number of predictor variables taken at random and k trees to be formed in order to obtain optimal results. According to Breiman (1996), the recommended value of k to be used in the bagging method that is tried is k = 50, it has given satisfactory results for the classification problem. According to Breiman and Cutler (2003), there are three ways formula to get the value of m to observe OOB errors:

$$m = \frac{1}{2}\left|\sqrt{p}\right|$$

$$m = \left|\sqrt{p}\right|$$

$$m = 2\,x\,\left|\sqrt{p}\right|$$

Where p is total variabel.

OOB data is used not to construct the tree, but to validate data on the corresponding tree. The random forest's misclassification value is suspected based on the OOB error generated by [23], which makes predictions on each OOB data in the relevant tree. Then, on average, around 36 percent of the original data cluster's observations, or one-third of the many trees created, will be OOB data. As a result, each of the initial data cluster observations is expected to account for around a third of the total number of trees in step 1. If is an observation from the original data cluster, then the random forest prediction result for each time becomes OOB data.

In a random forest, the OOB error is determined by the correlation between trees and the strength of each tree, with increasing the correlation increasing the OOB error and increasing the tree strength decreasing the OOB error [23]. The amount of misclassification of random forest prediction outcomes from all observations of the original data cluster is used to determine OOB error. Using a large number of trees, such as 1000 or more, according to Breiman and Cutler (2003), produces a more stable variable importance.

3 Problem Solution

3.1. GWR MODEL

Spatial effect testing was conducted to find out if there was a location effect on the research data. The hypothesis to be tested is $H_0 : I = I_0$ (no spatial correlation) vs $H_1 : I \neq I_0$ (there is a spatial correlation). Based on the results of Moran's I test shown in Table1, it was obtained that among all of the variables, Horizon Curvature (Kh) to Elevation (Elev) have spatial autocorrelation with a confidence interval of 95%. Because the variables in the study contain spatial autocorrelation, the model would be better off using geographically weighted regression model rather than using global regression model. The GWR model can contain spatial relationships because it contains spatial weights in it, so this model is more appropriate to use on data that has spatial autocorrelation.

Table 1. Spatial Autocorrelation Testing

Variable	P-value	Statistik Moran I	Result
Kh	1.43×10^{-10}	0.4929	Reject H0
Kv	3.23×10^{-8}	0.4947	Reject H0
H	4.35×10^{-9}	0.4952	Reject H0
K	4.39×10^{-8}	0.4830	Reject H0
M	1.04×10^{-9}	0.4885	Reject H0
E	1.925×10^{-9}	0.4916	Reject H0
Kmin	3.27×10^{-14}	0.4902	Reject H0
Kmax	1.98×10^{-8}	0.4963	Reject H0
Ka	1.41×10^{-15}	0.4942	Reject H0
Kr	2.20×10^{-16}	0.4908	Reject H0
Khe	2.63×10^{-8}	0.4938	Reject H0
Kve	9.35×10^{-12}	0.4881	Reject H0
S	1.52×10^{-8}	0.4907	Reject H0
An	1.28×10^{-7}	0.4876	Reject H0
Ae	7.08×10^{-8}	0.4968	Reject H0
Elev	5.82×10^{-8}	0.4818	Reject H0

Testing of GWR model parameters is simultaneously carried out to determine the effect of weighting in the process ofugaing soil texture parameters. The results of the parameter estimation test are simultaneously presented in Table 2. This test uses F test statistics based on the hypothesis H_0 : $\beta_j(u_i, v_i) = \beta_j$, where j =1, 2, .. k, vs H_1 : There is at least one $\beta_j(u_i, v_i)$ related to the location (u_i, v_i).

Based on Table 2, the statistical value of the F test on each model of soil particles proved to have a significant effect with a confidence level of 95%, thus it can be said that the weighting of the GWR model has a real effect on the resulting model and simultaneously all existing variables have a significant effect on the determination of soil texture. The next step is to perform a partial parameter estimation test, this is used to determine the effect of each variable on the variable of soil texture. Partial parameter presumption testing is based on hypotheses: $H_0 : \beta_j(u_i,v_i) = 0$ vs $H_1 : \beta_j(u_i,v_i) \neq 0$; $j = 1,2,\dots,k$. Reject H_0 if $|t\ test| > t_{(0.025,34)} = 2{,}03451$. Table 2 is a parameter test for area 1. By comparing the t test statistics with the critical point $t_{(0.025,34)}$, it was concluded that in the GWR model significantly affects the levels of sand texture, namely variable Ae.

In the silt, variables that have significant effect are variable H. Interpretation of the model at other area can be done in the same way. The GWR model at the level of sand texture based on Table 2 can be written as follows:

$\hat{Y}_1 =$ 1818944,822 - 2456849,895Kh + 6741920,337Kv - 9263712,927H - 342116,183K - 2656445,349M - 643,7964979E + 5106615,722Kmin - 126711,6561Kmax - 89558,63665Ka - 72,66770906Kr + 8544072,105Khe - 653577,4227Kve + 0,146746323S + 0,146746323An -6,055004494Ae - 0,008169045elv

The coefficient of determination (R^2) in the GWR model, texture of sand shows a value of 0.357. This means that the effect of predictor change (Kh to Elev) on the percentage of diversity of sand particle levels in area 1 by 35.7% and the other 64.3% is a large effect of other factors not described in the model.

Table 3 presents the GWR model based on the cut off of Random Forest. From the model, it can be seen that the value of the coefficient of determination (R^2) decreases in the sand and clay model, while in the silt model it increases. The RMSE value for all models on sand, clay and silt has increased. This indicates that the cut off model is not better than the GWR model with no cut off. This is supported by a partial test on each variable that is not significant. However, simultaneous testing is still significant in all the resulting models.

3.2. RF Model

Random Forest allows us to assess predictors' global, local, and partial effects on the spatial distribution of soil attributes. As a result, several interpretations of soil formation and the influence of certain features on the model can be deduced [26]. The outcome of this method is indicated in Table 3 below.

Table 2. Results for the prediction of sand, silt and clay using random forest models

	R^2	MSE	RMSE	bias
Sand	0.7897	23.51028	4.8487	-0.054147
Silt	0.7844	20.12793	4.48642	-0.086769
Clay	0.7870	12.32285	3.51039	-0.073716

In this Random Forest analysis process, the number of variables were randomized 5 times and the number of trees was 1000. The variation in the coefficients of determination was used to compare the variability of the prediction models for each PSF. This RF appears to be preferable to the GWR models. The superior performance of the RF model was shown in all PSF variable. As reported by Saraiva Koenow Pinheiro et al., (2018), tree models were thought to be an effective method for classifying a dataset into homogenous groups since they provided discrete output values in the terminal nodes (leaves).

Table 3. Estimated Parameters of GWR Model

	Sand		**Clay**		**Silt**	
Variable	Coefficient	t test	Coefficient	t test	Coefficient	t test
Intercep	1818944,822	0.693818	-670709.4394	-0.31078	-3246049.259	-1.89342
Kh	-2456849.895	-0.44914	-1031820.866	-0.24903	1112238.329	0.309523
Kv	6741920.337	1.295018	-3771791.837	-0.94961	-4209196.957	-1.22979
H	-9263712.927	-1.5107	4577866.329	0.953241	8672597.208	2.159678 *
K	342116.183	0.816757	-169317.0203	-0.51809	-171351.3689	-0.62329
M	-2656445.349	-0.37332	1298938.026	0.219133	-3420514.496	-0.73547
E	-643.7964979	-0.03256	7095.189643	0.46724	-5137.90597	-0.39587
Kmin	5106615.722	0.87436	-2007192.878	-0.41647	-7028305.887	-1.84081
Kmax	-126711.6561	-0.02124	2231535.613	0.481903	1453012.828	0.37107
Ka	-89558.63665	-0.43937	3597.878183	0.023235	79733.69435	0.59524
Kr	-72.66770906	-2.02281	44.20308822	1.660283	17.76190969	0.751561
Khe	8544072.105	1.550063	-4135171.855	-0.9618	-5194031.247	-1.43523
Kve	-653577.4227	-0.13698	-1403374.448	-0.38007	133242.2666	0.042535
S	0.146746323	1.22532	-0.03557148	-0.40062	-0.092935764	-1.17963
An	2.375778831	0.908506	-2.499221633	-1.27335	0.115596589	0.06725
Ae	-6.055004494	-2.12203 *	3.795642158	1.634822	0.749384797	0.400564
Elev	-0.008169045	-0.87356	0.013180686	1.846901	-0.006137487	-0.99835
	R^2= 0.357		R^2= 0.552		R^2= 0.376	
	RMSE = 8.4318		RMSE = 5.4022		RMSE = 5.5614	
	F test = 135.380*		F test = 92.358*		F test = 43.337*	

Significant if t test > t(0,025;33)= 2.03451

*) significant at level 5%

BASED ON CUT OF RANDOM FOREST (upper 0%IncMSE)

Table 3. Estimated Parameters of GWR Model

CLAY			**SAND**			**SILT**		
VAR	Coefficient	T-test	Var	Coefficient	T-test	Var	Coefficient	T-test
INTERCEPT	23.226	21.167	Intercept	49.585	32.704	Intercept	27.590	19.916
KH	14731.853	1.043	Kh	-10454.300	-0.446	Kh	-16350.929	-0.797
H	-541.636	-0.075	H	-18923.510	-1.096	K	-1.140	-0.466
K	-1.983	-1.078	K	0.562	0.260	E	19.095	0.411
M	-393.281	-0.074	E	21.473	0.541	Kmin	-6831.935	-0.472
E	5.932	0.175	Kmin	31257.804	1.426	Kmax	17822.041	1.307
KMAX	-10074.849	-0.834	Kmax	2220.130	0.138	Ka	0.232	0.091
KA	1.807	0.902	Khe	11403.659	1.177	Khe	-4148.004	-0.471
KVE	12293.446	1.042	Kve	9996.396	0.572	Kve	-20515.687	-1.308
			Elev	-2.529	-1.504	Elev	0.543	0.339
	R^2= 0.194			R^2= 0.216			R^2= 0.414	
	RMSE = 29.21184			RMSE = 25.112			RMSE = 9.796	
	F test = 46.623*			F test = 121.47*			F test = 109.969*	

SIGNIFICANT IF T TEST > T (0,025;41)= 2.0195 ; T (0,025;40)= 2.02107
***)SIGNIFICANT AT LEVEL 5%**

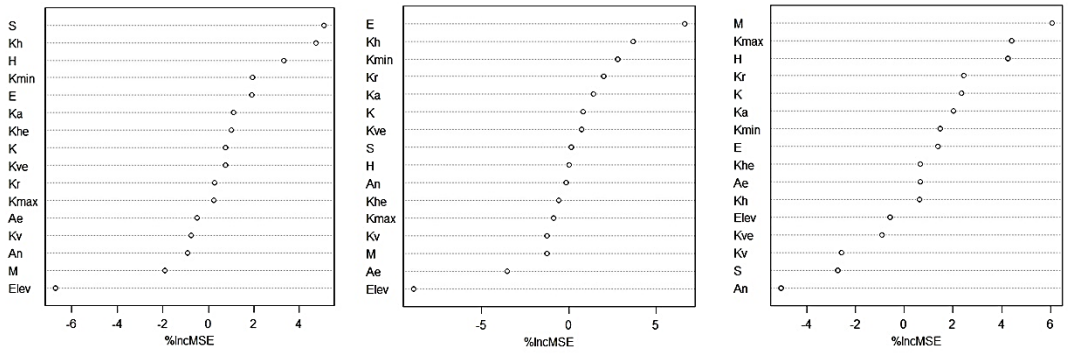

Fig. 3: Importance of the environmental covariates derived from the RF models for sand, silt and clay.

The importance of the environmental covariates in each tested RF model is shown in Fig. 3. Based on the investigated attributes, the results revealed various covariate combinations. Slope is the most importance factor to estimate sand, elevation is for silt and unsphericity curvature is for clay. This variable explained was considered moderately satisfactory for sand (19.08%), clay (26.11%) and silt (19.78%). The considered characteristic, as well as other factors, influence the importance of the variables predicted by the RFs [14].

The results obtained are not superior to those achieved in previous studies [15], [16] but still comparable. The number of samples used and the selected independent variables can cause differences in the results of this study with previous studies. Nonetheless, a different predictor variable was utilized in the previous report than in this study. However, the low performance of the result in this study was due to the source material's small-scale variation and relative erosion/deposition along the slope, which could not be captured by the covariates' 30 m spatial resolution. That's problems also found in former study [17].

This PSF soil looks more suitable to be modelled using RF compared to GWR. This is because GWR which has the basic principle that "everything is related to everything else, but near things are more related than distant things" [18] does not match the nature of PSF. PSF is strongly influenced by topography not by the proximity of each soil particle. Parent material has a considerable impact on the PSF, encouraging an early pedogenic process and structural development [1]. As a result, RF should be considered a feasible alternative for modelling PSFs. This is compounded by the fact that RF models provide an estimate of the relative importance of the variables in the model [19]. This backed up the statement that using random forest models to predict PSF content in soils provided good performance [15].

4. Conclusion

GWR and RF were used to create prediction models for sand, silt, and clay, and all variables passed the statistical assumption test. In terms of prediction, the RF model outperforms the GWR model, and it should still be considered a viable alternative for modelling PSFs. This research suggests a link between topsoil qualities and terrain attributes, which may be determined using field observations and model predictions. More study is needed to develop more efficient input factors that will aid in the precision of soil variability and the accuracy of soil map products.

References:

[1] H. Saraiva Koenow Pinheiro, W. de Carvalho Junior, C. da Silva Chagas, L. Helena Cunha dos Anjos, and P. Ray Owens, "Prediction of Topsoil Texture Through Regression Trees and Multiple Linear Regressions," *Artic. Rev Bras Cienc Solo*, vol. 42, p. 170167, 2018.

[2] C. Radim and N. Zdenka, "Study of input parameters of layered half-space used for soil modelling," *WSEAS Trans. Appl. Theor. Mech.*, vol. 15, pp. 194–205, 2020.

[3] A. B. Mcbratney, *Pedometrics*. Cham: Springer International Publishing, 2018.

[4] A. Moufakkir, A. Samaouali, A. Elbouzidi, E. A. Salah, and A. Dinane, "The Influence of the Percentage of Porosity on the Thermal

Conductivity of a Composite Material, for Example Clay," *WSEAS Trans. Environ. Dev.*, vol. 16, pp. 566–572, Jun. 2020.

[5] A. Gobin, P. Campling, and J. Feyen, "Soil-landscape modelling to quantify spatial variability of soil texture," *Phys. Chem. Earth, Part B Hydrol. Ocean. Atmos.*, 2001.

[6] K. Liao, S. Xu, J. Wu, and Q. Zhu, "Spatial estimation of surface soil texture using remote sensing data Spatial estimation of surface soil texture using remote sensing data," 2013.

[7] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically Weighted Regression," *J. R. Stat. Soc. Ser. D (The Stat.*, vol. 47, no. 3, pp. 431–443, 1998.

[8] D. C. Wheeler, "Geographically weighted regression," in *Handbook of Regional Science*, 2014.

[9] M. Fischer and A. Getis, *Handbook of Applied Spatial Analysis*. New York: Springer, 2010.

[10] G. Biau, "Analysis of a random forests model," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1063–1095, 2012.

[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[12] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[13] L. Breiman and A. Cutler, "Manual for Setting Up," *Using, Underst. Random For.*, vol. 4, 2003.

[14] S. I. C. Akpa, I. O. A. Odeh, T. F. A. Bishop, and A. E. Hartemink, "Digital Mapping of Soil Particle-Size Fractions for Nigeria," *Soil Sci. Soc. Am. J.*, vol. 78, no. 6, pp. 1953–1966, 2014.

[15] C. da S. Chagas, W. de Carvalho Junior, S. B. Bhering, and B. Calderano Filho, "Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions," *CATENA*, vol. 139, pp. 232–240, Apr. 2016.

[16] M. Ließ, B. Glaser, and B. Huwe, "Uncertainty in the spatial prediction of soil texture," *Geoderma*, vol. 170, pp. 70–79, Jan. 2012.

[17] K. Vaysse and P. Lagacherie, "Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France)," *Geoderma Reg.*, vol. 4, pp. 20–30, 2015.

[18] T. M. Oshan, Z. Li, W. Kang, L. J. Wolf, and A. Stewart Fotheringham, "MGWR: A python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale," *ISPRS Int. J. Geo-Information*, vol. 8, no. 6, 2019.

[19] D. R. Cutler *et al.*, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Henny Pramoedyo has Conceived and designed the analysis

Sativandi Riza has implemented the Random Forest Algorithm in Rstudio

Novi Nur Aini has implemented the GWR algorithm in Rstudio

Danang Ariyanto has responsible for Statistics

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

The authors are grateful to the Department of Mathematics and Natural Science, University of Brawijaya which is supported to this study by providing the support and funding.